When Kahneman meets Manski: Using dual systems of reasoning to interpret subjective expectations of equity returns

Fabian Gouret, Guillaume Hollard

When Kahneman meets Manski: Using dual systems of reasoning to interpret subjective expectations of equity returns*

Fabian Gouret[†], Guillaume Hollard[‡]

October 11, 2010

Abstract

To understand how decisions to invest in stocks are taken, economists need to elicit expectations regarding risk-return trade-off. One of the few surveys which has elicited such expectations is the Survey of Economic Expectations in 1999-2001. Using the data from this survey, Dominitz and Manski find considerable heterogeneity across respondents that cannot be explained by simple models of expectations formation. Adapting a principle of dual-reasoning borrowed from Kahneman, this paper classifies respondents according to their sensitivity to some pathologies. We find a substantial amount of unobserved heterogeneity between the least and the most sensitive respondents. We then sketch a model of expectations formation.

JEL: C42, C81, D84, G1, G23

Keywords: Subjective probability distribution, dual system of reasoning, investment funds.

*We thank the Editors Charles Bellemare and Charles Manski, as well as two anonymous referees for valuable and detailed criticisms. We are also grateful to Ibrahim Ahamada, Adeline Bachellerie, Thomas Barré, Andrew Caplin, Jézabel Couppey, Emmanuel Flachaire, Guy Lacroix, Andrew Schotter, Jean-Christophe Vergnaud and participants at the Workshop on Subjective Beliefs in Econometric Models at Laval University and in seminars at Universitat de Barcelona, University of Paris 1 and Cergy University for helpful discussions. In addition, we thank Jordi Roca-i-Solanelles for his kind assistance with the tables and figures during the revision process. Fabian Gouret acknowledges the support of the Barcelona Graduate School of Economics and of the Government of Catalonia, as well as the financial support from the Spanish Ministry of Science and Technology provided by a Juan de la Cierva contract and via the project CICYT ECO2008-04997.

†Corresponding author. Departament de Teoria Econòmica and CAEPS, Universitat de Barcelona, Diagonal 690, 08034 Barcelona, Spain (Email: `fabiangouret@ub.edu`).

‡Paris School of Economics and CNRS, 106-112 Bd de l'Hôpital, 75647 Paris Cedex 13, France (Email: `guillaume.hollard@univ-paris1.fr`).

1 Introduction

The decision to invest in stocks requires an assessment of the risk-return trade-off. To measure this arbitrage, financial researchers should be interested in eliciting potential investors' expectations regarding equity returns. Yet the prevalent practice is to use realized returns as a proxy for expected returns. By doing so, researchers neglect some fundamental issues in finance such as the degree of homogeneity between agents' expectations and their types of expectations. These issues will remain unexplored if we do not ask potential investors what they expected ex-ante (Bossaerts, 2002). Various surveys have included questions about individuals' expectations over equity returns. But most of these elicit point forecasts or Yes/No predictions which tell us nothing about respondents' perceived risk.[1] To provide an empirical basis for the study of this risk-return trade-off, we need to elicit subjective probability distributions. But only few surveys to date have included such questions.

Two of the recent exceptions are Dominitz and Manski (2010, 2007): the reserarch here is based on surveys that measure probabilistically the beliefs that Americans hold about mutual-fund returns. In particular, Dominitz and Manski (2010) -hereafter DM (2010)- present an analysis of answers to expectations questions in the Survey of Economic Expectations (SEE) that took place in three waves over the period July 1999-March 2001. They first pose the following scenario (hereafter the SEE scenario):

> Please think about the type of mutual fund known as a diversified stock fund. This type of mutual fund holds stock in many different companies engaged in a wide variety of business activities. Suppose that tomorrow someone were to invest one thousand dollars in such a mutual fund. Please think about how much money this investment would be worth one year from now.

[1] An example of a question asking for a point forecast is the following from the Michigan Survey of Consumers (MSC): "[...] What is the annual rate of return that you would expect a broadly diversified portfolio of U.S. stocks to earn, on average, over the next three years?". An example of a question eliciting Yes/No predictions is also provided by the MSC: "Would you expect the average return over the next ten to twenty years to be much different than this? Yes/No".

They then ask respondents to answer various questions regarding their expectations were they to face this scenario. In particular some expectations questions take the following form (hereafter the *probabilistic questions*):

> What do you think is the PERCENT CHANCE (or CHANCES OUT OF 100) that, one year from now, this investment would be worth over R?

Respondents face a sequence of such probabilistic questions posed for four investment thresholds. The thresholds are determined by the respondent's answer to two preliminary questions according to an algorithm detailed in DM (2010) (hereafter the *preliminary questions*):

> What do you think is the LOWEST amount that this investment of $1000 would possibly be worth one year from now?
>
> What do you think is the HIGHEST amount that this $1000 investment would possibly be worth one year from now?

These preliminary questions are thought to decrease overconfidence with respect to central tendencies and anchoring problems wherein respondents' beliefs are influenced by the questions posed (Manski, 2004). The answers to these preliminary questions are not analyzed in DM (2010). As we will explain, a key feature of our paper is to construct a setting that allows us to extract useful information from these questions (a similar idea is developed in Dominitz and Manski (1997, p.860)).

DM (2010) consider the sequence of probabilistic questions; they assume that each respondent i has a normal subjective probability distribution of the one-year-ahead return, $\mathcal{N}(\mu_i, \sigma_i^2)$, which they estimate (hereafter called the *fitted* subjective distribution). They find considerable heterogeneity in reported beliefs in the sense that the values of (μ, σ) of the fitted subjective distributions vary considerably across individuals. Given that respondents may differ in the way that they use public information to form their expectations, DM attempt to identify types of expectation-formation models: people might think that stock markets follow random walks, display persistence or are characterized by mean-reversal. But even a mixture of such models of expectations formation accounts for only 4% of the variation in respondents' answers. This unexplained heterogeneity leads DM (p.22) to conclude that the remaining question is "*why these processes vary so much across persons*", and that there is no "*parsimonious specification of types* (that)*can adequately explain the diverse expectations of the Michigan and SEE respondents.*"

In the light of this apparent lack of regularity in stated expectations, a skeptical economist might conclude that subjective expectations data are not useful. There are indeed several pathologies which are likely to affect the quality of the data: probabilistic questions are not easy for ordinary people, financial literacy is often lacking in the general public and framing effects are known to affect survey responses. Even if the preliminary questions help respondents to answer the probabilistic questions, some of these pathologies are likely to survive.

This paper takes these critics seriously and considers that these pathologies are an important source of heterogeneity. Suppose that some respondents have a hard time answering probabilistic questions and thus rely on rules of thumb which are prone to biases. They are then likely to provide nothing but noise, i.e. barely interpretable data. However, a fraction of the respondents may behave roughly as expected and reveal valuable information: they take the question seriously and, after some serious reflection, produce their "true" expectation. To see if such heterogeneity is at work, we need to understand how respondents cope with probabilistic questions and classify them according to the type of reasoning used. To do so, we adapt a principle of dual reasoning borrowed from Kahneman (2003). In a nutshell, we thus revisit Manski's approach in the light of Kahneman (and Tversky).

The challenge is to operationalize this intuition. Section 2 explains how a dual system of reasoning can accurately describe the traditional pathologies that affect survey data, and how the specific design of the SEE allows us to build a proxy for this dual system. We consider the lowest and highest values elicited in the preliminary questions as a confidence interval. Under some assumptions, we are able to calibrate a probability distribution based on the preliminary questions only (hereafter called the *calibrated* probability distribution). This calibrated probability distribution can provide a prediction of the answers to probabilistic questions. The distance between a predicted value and the actual value that the respondent provides to a probabilistic question is then a measure of coherence. This allows us to classify individuals according to their level of coherence.

In Section 3, we analyze the cross-sectional distribution of the values of (μ, σ) of the fitted subjective distributions conditional on this measure of coherence. We find a highly significant positive monotonic relationship between expected returns μ and perceived risks σ for the most coherent respondents. This result is in sharp contrast to the same cross-sectional distribution for the least coherent which exhibits no such pattern, nor any other particular relationship. As anticipated by our cognitive approach, it is as if the answers of the least coherent respondents are essentially nothing but noise. In line with our approach, we also find that our measure of coherence is related to education and income, i.e. individual attributes known to be related to cognitive ability.

Section 4 mainly focuses on the most coherent individuals. We first show that a simple linear regression of μ on σ explains a substantial part of the total variation in μ for these respondents. To interpret this surprising pattern, we suggest that the most coherent respondents have different portfolios in mind, ranging from investments with small expected returns and low risk to investments with larger expected returns but greater risk. At the extreme, those who consider a safe portfolio should expect the portfolio to perform like a risk-free asset. With this particular interpretation in mind, we view the intercept and the slope coefficients of the simple linear regression of μ on σ as the risk-free rate and the ex-ante price of risk, respectively. This estimated price of risk is weakly pro-cyclical, as it is statistically but not substantially lower in the last wave, a period characterized by a steady drop in the stock market. With only three waves the time-series dimension is however too short to consider this result as robust. Last, we provide more speculative results on the nature of expectations formation. Section 5 concludes.

2 Survey pathologies and dual systems of reasoning

Subsection 2.1 below reviews the objections addressed to subjective survey data in economics. We claim in Subsection 2.2 that these objections can usefully be addressed using Kahneman (2003)'s dual system of reasoning. Subsection 2.3 then explains how the particular format of the SEE survey allows us to construct some measures of coherence which account for individual sensitivity to the cognitive biases occurring in such surveys.

2.1 The case against subjective data

Economists are often skeptical of subjective data, in particular expectations data (Manski, 2004, p.1337). Economists usually prefer to infer expectations from data on observed choices. The roots of this skepticism are not easy to track down. We have identified four kinds of objections that can be addressed to probabilistic questions:

i. Some individuals are sensitive to framing effects. Following Tversky and Kahneman (1974), a large literature has shown that minor changes to the framing of a problem may lead to unanticipated changes in answers, especially if respondents are not well informed (e.g., Hurd, 1999). This is a problem as survey questions are supposed to elicit "true" values, i.e. values which are independent of the elicitation technique used.

ii. Some individuals have trouble thinking in probabilistic terms. If the question forces them to give numerical probabilities, then the question "*forces them to operate in a "mode" which requires "more mental effort" and is therefore more prone to interference with biasing tendencies*" (Zimmer, 1984, p.123).

iii. Some individuals lack some relevant knowledge. Van Rooij *et al.* (2007) provide evidence that a majority of Americans have only limited knowledge of bonds and stocks, the concept of risk diversification, and the working of financial markets. Respondents with less financial literacy are more likely to be prone to biases (e.g., the thresholds of the probabilistic questions can influence their probabilistic answers). If this is the case, respondents with less financial literacy will likely provide noisier answers.[2]

iv. Last, a recurrent objection is that some individuals do not take the questions seriously. Answering questions, particularly numerical ones, represents a cognitive burden and there is no incentive to provide such effort. Some individuals may thus provide automatic answers, without putting in the required effort.

Each of these four objections has similar consequences: answers to survey questions are generated by some heuristics which lead to cognitive biases. If *all* respondents are strongly affected by at least one of these pathologies, there are serious reasons to doubt the validity of subjective probabilistic expectations. But if these pathologies affect some respondents more than others, the quality of collected data may greatly vary across individuals. We thus face a problem of unobserved heterogeneity regarding the individual sensitivity to these pathologies (Flachaire and Hollard (2008) provide empirical support in the context of value elicitation). Controlling for such heterogeneity requires opening the black box of the cognitive process used by each survey respondent. This sounds like a difficult task. The next subsection thus explores the possibility that the myriad individual cognitive processes can be classified into two broad types which make sense regarding the quality of the answers provided.

2.2 Dual systems of reasoning

A dual system of reasoning consists of two systems which can be used to perform a cognitive task. Several such models exist in psychology. Roughly speaking, they are all based on a distinction between two systems: one is usually associated with intuition and the other with reasoning. This Subsection builds on a dual system of reasoning model described by Kahneman

[2]Nonetheless, note that basing the thresholds on the preliminary questions asking the respondents for the lowest and highest possible future values of the investment allows to reduce the “anchoring” problem, as argued by Dominitz and Manski (1997, p.858) and Manski (2004, p.1347).

(2003) in his Nobel lecture. In particular, the two different systems will be labeled System 1 and System 2. System 1 is usually associated with intuition. It is fast, automatic and proposes intuitive answers to judgment problems as they arise. System 2, in contrast, is controlled and encompasses analytical intelligence.[3] System 2 requires effort. Note that the term "System" is only "*a label for collections of cognitive processes that can be distinguished by their speed, their controllability, and the contents on which they operate*" (Kahneman and Frederick, 2005, p.267).

We now turn to the specific task proposed in the SEE survey, namely a sequence of probabilistic questions. Both systems are able to provide answers to these. System 1 will provide fast and easy answers, such as round numbers or focal values. System 2 will provide more analytical answers which entail effort. In contrast, the answers from System 1 will be noisier, and are likely affected by minor details in the survey design that are difficult to control for.

The analysis presented in this paper is based on the idea that the identified pathologies affecting survey responses result from respondents using System 1 rather than System 2. We now reconsider the four objections listed in Subsection 2.1 in the light of a dual system of reasoning. (i.) Framing effects occur when System 2 only lightly monitors judgement, as noted by Kahneman (p.1467). Individuals thus provide automatic answers without even realizing that they are sensitive to biases. (ii.) Thinking in probabilistic terms is directly linked to System 2 which encompasses analytic intelligence. Some shortcuts can be used, i.e. heuristics, but they may well lead to erroneous answers. (iii.) Now consider individuals who lack relevant knowledge. They are likely to be uncertain about their answers. Thus, any numerical cue may influence them, in particular if that cue is plausible. They might know how to handle probabilities, but lack the relevant knowledge to perform calculations. (iv.) Those who do not

[3] A more complete description of the dual system of reasoning also recognizes that System 2 plays a monitoring role. It is supposed to override System 1 when some erroneous decisions are made. Errors of intuitive judgement occur when System 2 does not monitor judgement. Kahneman (p.1467) emphasizes that "*the prevalence of framing effects, and other indications of superficial processing* [...] *suggest that System 2 monitors judgements quite lightly.*"

care enough about the survey or who refuse to put much effort into their answers will use the effortless system, i.e. System 1.

Thus, while these four pathologies are all different, they do have a common element: they consist of the use of cognitive processes belonging to System 1. We are thus facing a heterogenous population, composed of "System 1" and "System 2" respondents. As a direct consequence, if we can separate "System 1" from "System 2" respondents, we can then isolate corrupted answers from higher-quality data.

2.3 Building a proxy of the dual system approach

As already mentioned, the survey analyzed in this paper uses a particular design. The two preliminary questions ask only for the lowest and highest possible future values of the investment. These questions are fairly easy and do not imply any probabilistic calculations, in contrast to the sequence of probabilistic questions asking respondents to provide a sequence of points on their subjective complementary cumulative distribution function. Providing *coherent* answers between the preliminary questions and the sequence of probabilistic questions is not easy, and respondents may well be prone to the pathologies described above. Our claim is that only those who use System 2 to answer these questions will provide coherent answers, i.e. they use the same underlying probability distribution to answer both types of question. We will thus use a measure of coherence as a proxy for the reasoning system that individuals use.

The preliminary questions allow us to calibrate a probability distribution which will serve to predict the answer to a probabilistic question. The distance between the predicted and the actual answers yields a proxy for the reasoning system.

The preliminary questions and the calibrated probability distribution. The two preliminary questions allow us to calibrate a probability distribution under a number of assumptions. The elicited lowest and highest possible future values of the investment, denoted $R_{i,min}$ and $R_{i,max}$, are not interpreted literally as minimum and maximum values. Following

Dominitz and Manski (1997), the phrases "lowest possible" and "highest possible" are too vague to justify this formal interpretation. Instead, $R_{i,min}$ and $R_{i,max}$ reflect the general region of respondent i's subjective distribution of $\widetilde{R}_i$, the one-year-ahead investment value based on these two preliminary questions. To construct a measure of coherence, we first make the two following assumptions which will allow us to calculate the parameters of the calibrated probability distribution.

Assumption 1 $Prob\left(\widetilde{R}_i < R_{i,min}\right) = Prob\left(\widetilde{R}_i > R_{i,max}\right) = \alpha, \forall i$

Assumption 1 says that the probability that $\widetilde{R}_i$ lies within the interval $[R_{i,min}, R_{i,max}]$ is $(1-2\alpha)$, and the remaining part of the distribution, 2α, is equally allocated to the left and the right of the distribution.

Assumption 2 $\widetilde{R}_i \sim \mathcal{N}(\widetilde{\mu}_i, \widetilde{\sigma}_i^2)$: *the calibrated probability distribution is a normal distribution with mean* $\widetilde{\mu}_i$ *and variance* $\widetilde{\sigma}_i^2$.

Three observations are in order concerning Assumptions 1-2. First, Assumption 1 says that the interval is the same for all the respondents, or, in other words, α is not individual-specific. This assumption may appear to be strong, and we will discuss at the end of this Section and in the online appendix one way in which to relax it. Second, Assumption 2 imposes that $\widetilde{R}_i$ is normally distributed. The assumption of normality is crucial to calibrate a distribution based on the preliminary questions. This assumption is debatable, even if it is standard in finance and is also used by DM. But it appears as the simplest alternative, as we need to calibrate a distribution using only $R_{i,min}$ and $R_{i,max}$. Third, it is reasonable to think that α should be small. But the notion of "small" is admittedly arbitrary. We will carry out the empirical analysis in the main text assuming that $\alpha = 0.01$, and will present a set of robustness checks with alternative values for α in the online appendix.

A normal distribution is symmetric around its mean, so under Assumptions 1-2 the mean of the calibrated normal distribution is $\widetilde{\mu}_i = \frac{R_{i,min}+R_{i,max}}{2}$. Regarding the standard deviation, we have from Assumption 1 that Prob $\left(\widetilde{R}_i < R_{i,min}\right) = \alpha$. Let $\Phi(.)$ denote the standard normal cumulative distribution function. The standard deviation is then equal to $\widetilde{\sigma}_i = \frac{R_{i,min}-\widetilde{\mu}_i}{\Phi^{-1}(\alpha)}$. $\widetilde{R}_i$ thus follows the calibrated normal distribution:

$$\widetilde{R}_i \sim \mathcal{N}\left(\frac{R_{i,min}+R_{i,max}}{2}, \left(\frac{R_{i,min}-\widetilde{\mu}_i}{\Phi^{-1}(\alpha)}\right)^2\right) \tag{1}$$

This calibrated probability distribution provides a predicted answer to a probabilistic question. Note that if $R_{i,min} = R_{i,max}$, $\widetilde{R}_i$ has a degenerate distribution with all of the probability mass at the single point $R_{i,min}$: in this case $\widetilde{\mu}_i = R_{i,min}$ and $\widetilde{\sigma}_i = 0$, $\forall\alpha$.

The probabilistic questions. The sequence of probabilistic questions elicits the percentage chance that the one-year ahead investment value will be worth more than four different thresholds (posed in increasing order). Hence, for each respondent i, we observe $Q_{i,k} \equiv \text{Prob}\left(R_i > R_{i,k}\right)$, $k = 1, 2, 3, 4$, where R_i denotes the one-year-ahead investment value, and $R_{i,1} < R_{i,2} < R_{i,3} < R_{i,4}$ are the four investment value thresholds about which the respondent is queried. The probabilistic expectations data are used to find the value of (μ_i, σ_i) of each fitted subjective distribution. Let $F(R; \mu_i, \sigma_i)$ denote the cumulative normal distribution function with mean μ_i and standard deviation σ_i, evaluated at any point R. For each respondent i, we find the (μ_i, σ_i) that solves the following least-squares problem:

$$\inf_{\mu_i,\sigma_i} \sum_{k=1}^{4} [(1 - Q_{i,k}) - F(R_{i,k}; \mu_i, \sigma_i)]^2 \tag{2}$$

We thus proceed as in DM to find the values of (μ, σ).[4] Remember that we here analyze the cross-section distribution of the values of (μ, σ) conditional on a measure of coherence. The value of a calibrated probability distribution $(\widetilde{\mu}_i, \widetilde{\sigma}_i)$ only serves to compute the measure of coherence for respondent i, as described below.

[4]For more details, see Dominitz (2001, Appendix A) who fits (log-normal) person-specific subjective income distributions.

A measure of coherence. We now have *two* measures of the one-year-ahead investment value: R_i and $\widetilde{R}_i$. The first, R_i, comes from the sequence of probabilistic questions, while $\widetilde{R}_i$ is derived from the two preliminary questions. Given that we now have the parameters $(\widetilde{\mu}_i, \widetilde{\sigma}_i)$ of the calibrated probability distribution of $\widetilde{R}_i$, we can compute $\text{Prob}\left(\widetilde{R}_i > R_{i,k}\right)$: this is a prediction of $Q_{i,k}$. These two probabilities are more likely to differ when the respondent does not use System 2. We thus classify individuals according to the gap between $\text{Prob}\left(\widetilde{R}_i > R_{i,k}\right)$ and $Q_{i,k}$. We consider the distance at the first threshold $R_{i,1}$:

$$d_i^{98} = \left| Q_{i,1} - \text{Prob}\left(\widetilde{R}_i > R_{i,1}\right) \right| \tag{3}$$

Note that $d_i^{98} \in [0, 1]$.[5] The smaller this distance, the more coherent is the respondent. The superscript "98" indicates that the parameters of the calibrated probability distribution have been computed assuming that $\alpha = 0.01$ (so that $\text{Prob}\left(\widetilde{R}_i \in [R_{i,min}, R_{i,max}]\right) = 0.98$).

Alternative measures of coherence. Instead of considering d_i^{98}, we could have compared (μ_i, σ_i) with $(\widetilde{\mu}_i, \widetilde{\sigma}_i)$ to construct a measure of coherence. But note that d_i^{98} requires fewer assumptions. It does not require any estimation of the distribution parameters based on the four probabilistic questions, and so does *not* depend on the normality assumption to find the value of (μ_i, σ_i). It only depends on the assumptions made regarding the calibrated probability distribution. Also note that the distance d_i^{98} is a coherence criterion based on the first threshold value $R_{i,1}$. We could have computed the distance at any one of the other three thresholds. However, the first threshold is more natural because the answer for the first threshold is not influenced by the answers for the subsequent thresholds, while the reverse is not true. Once a respondent has answered $Q_{i,1}$, his answers to the next three thresholds must subsequently weakly fall ($Q_{i,1} \geq Q_{i,2} \geq Q_{i,3} \geq Q_{i,4}$).[6] Hence, we prefer d^{98} as a coherence measure and this

[5] The extreme case $d_i^{98} = 1$ can occur for a degenerate calibrated probability distribution, such that $R_{i,min} = R_{i,max} < R_{i,1}$ and $Q_{i,1} = 1$. Only one respondent interviewed in the SEE falls into this category. He answered $R_{i,min} = R_{i,max} = 120$ at the preliminary questions, implying a prediction $\text{Prob}\left(\widetilde{R}_i > R_{i,1}\right) = 0$. But at the first threshold $R_{i,1} (= 500)$ of the sequence of probabilistic questions, he answered $Q_{i,1} = 1$.

[6] Remark: if a probability elicited at threshold $R_{i,2}$, $R_{i,3}$ or $R_{i,4}$ was higher than one elicited previously, the

is what is analyzed in the paper below.

Even so, we have also carried out the analysis described in this paper with seven other measures of coherence, each of which has its own advantages compared to d^{98}. The robustness of our results are thus checked for three waves and eight different measures of coherence. Before we go further, we briefly describe the alternative measures considered.[7] First, and as noted above, obvious alternative measures of coherence are obtained by varying the parameter α ($\alpha = 0.05$ and $\alpha = 0.10$). These measures are called d^{90} and d^{80}. Second, despite our argument in favor of distances computed at the first threshold, it can be seen that one potential inconvenience is that the value $R_{i,1}$ can vary across respondents. Each respondent's sequence of probabilistic questions included a \$1000 threshold, and we can consider this invariance as being a desirable property. We have considered a measure of coherence, d_{1000}, computed at the \$1000 threshold for all respondents. Third, we could have used a measure based on the four thresholds. We therefore also carry out the analysis with a measure which averages the absolute deviations over the four thresholds ($\overline{d}$), as well as a Kolmogorov distance, i.e. the greatest of the four distances (d^{Kol}). Last, and as discussed above, we may consider Assumption 1 to be particularly strong, given that α is not individual-specific. We have taken a distinct approach by assuming that respondents are in fact coherent, but that the value of α varies across individuals. The value α_i is given by the value that minimizes the distance between Prob $\left(\widetilde{R}_i > R_{i,1}\right)$ and $Q_{i,1}$, and α_i thus becomes the measure of coherence for respondent i. Similarly, we have considered the distance at the \$1000 threshold: the individual-specific quantile which minimizes this distance is called $\alpha_{i,1000}$.[8]

interviewer informed the respondent and requested a replacement response (see Dominitz and Manski, 2004, pp.4-5, pp.8-9).

[7]More details concerning these measures of coherence are presented in Sections B and C of the online appendix. Section B explains precisely how we construct these measures and the corresponding results. Section C compares the rankings of respondents based on the different measures. The rankings are very similar, with the Spearman rank correlation coefficients between d^{98} and the alternative measures usually being over 80 percent, for all SEE waves. The two exceptions are the Spearman's rhos of d^{98} with $\overline{d}$ and d^{Kol} which are around 30 and 55 percent, respectively. The null hypothesis of statistical independence is however always rejected at the 1 percent significance level.

[8]Given that α_i and $\alpha_{i,1000}$ relax Assumption 1, they can appear as much better than d^{98} and the other alternative measures of coherence. The inconvenient of α_i and $\alpha_{i,1000}$ is that they are not uniquely defined for

3 Application

This Section analyzes the cross-sectional distribution of the values of (μ, σ) of the fitted subjective distributions conditional on our measure of coherence. Subsection 3.1 describes the sample used, and Subsection 3.2 checks whether our measure of coherence does indeed capture some unobserved heterogeneity across respondents. Last, Subsection 3.3 considers the relationship between our measure of coherence and observed individual characteristics.

3.1 The data

The data used in this paper are drawn from the three waves of the SEE where the questions on equity returns were asked: wave 12 (where interviews were conducted between July 1999 and November 1999), wave 13 (February 2000-May 2000) and wave 14 (September 2000-March 2001). 1651 respondents were interviewed in these three cross-sections. Of these 1651 respondents, 1284 answered the preliminary questions eliciting the lowest and the highest possible values of the investment one year ahead. Of these 1284 respondents, 1125 answered the sequence of questions posed for the four thresholds. We can compute the measure of coherence d^{98} for these 1125 respondents, but we cannot fit a subjective distribution for some of them. 120 respondents reported the same probability values at all four of the specified thresholds: the absence of variability in these responses makes it impossible to fit a subjective distribution for them. Of the 1005 remaining respondents, we drop five other respondents as all four of their elicited probabilities ($Q_{i,k}$, $k = 1, ..., 4$) take the values of 0 or 1 (these five cases correspond to the case where the responses are $(1, 1, 1, 0)$). For each of the 1000 remaining respondents, we solve the least-squares problem in Equation 2 to find the value of (μ_i, σ_i) for each respondent i. We had some difficulty in obtaining a stable and unique fitted value of (μ, σ) for 21 of the respondents, who we excluded. Hence, our final sample is composed of 979 respondents, implying an effective response rate of about 60% ($\simeq 979/1651$).

some respondents (see Subsections B2 and B4 of the online appendix).

3.2 Does the measure of coherence matter?

Table 1 describes the cross-section distribution of the values of (μ, σ). Note that, as in DM, the data are rescaled, such that μ_i denotes the expected return of respondent i (e.g., $\mu_i = 0.03$ means that the expected value of the investment one year from now is 1030).[9] The normality of subjective distributions allows us to completely describe each respondent by a single dot in a (σ, μ) space, as in Figure 1. Panels (a), (b) and (c) of this Figure summarize the cross-section distribution of the perceived risk and return of respondents interviewed in waves 12, 13 and 14, respectively.

Figure 1 clearly shows that there is a great deal of heterogeneity in beliefs across individuals, as noted by DM. No clear pattern in this heterogeneity emerges. For instance, it is not obvious that more risk corresponds to higher returns. At first glance, it might be thought that this expectations data does not contain much relevant information. We think that this apparent absence of an interpretable pattern results from the noise due to pathologies that affect survey responses. If this is the case, we expect the introduction of our measure of coherence to make a difference, given that it is a proxy for the type of cognitive reasoning used.

To evaluate the effect of our cognitive variable, we rank the N_t respondents according to the distance d^{98} for each wave t, beginning with the lowest d_i^{98} value, i.e. the most coherent respondent in wave t. We separate respondents into three equally-sized categories, $j = 1, 2, 3$: those with high levels of coherence ("Coherent 1"), those with low levels of coherence ("Coherent 3") and the intermediate ("Coherent 2"). The presence of this intermediate group will allow us to draw a sharper distinction in the cross-section distribution of (μ, σ) between the most and the least coherent. Table 2 summarizes the number of respondents included in each category for each wave.

Figure 2 presents the cross-section distribution of the values of (μ, σ) by category of coher-

[9] For the sake of comparison, Note (i.) at the bottom of the Table provides the cross-section distribution of the values of (μ, σ) in DM (Table 5, p.31). The summary statistics are similar to theirs, so the interested reader should read their paper for more description.

ence. Panel (a) presents a scatter diagram for the most coherent in wave 12 of the expected return against the risk, and Panel (c) the same scatter diagram for the least coherent in wave 12.[10] Panel (a) shows considerable heterogeneity in the cross-section distribution of (μ, σ) for the most coherent, but also suggests that the expected return is positively (and perhaps linearly) related to the risk for these respondents. This positive monotonic relationship is confirmed by a Spearman's rho equal to 0.60 and significantly different from zero at the 1 percent significance level.[11] On the contrary, Panel (c) shows that the cross-section distribution of the values of (μ, σ) for the least coherent exhibits no particular pattern: the Spearman correlation coefficient is 0.09 and statistically insignificant. The same finding pertains in waves 13 (Panels (d) and (f)) and 14 (Panels (g) and (i)): there is a highly statistically significant positive monotonic relationship between μ and σ for the most coherent, but no particular structure for the least coherent, as if the latter were only providing noise.

The difference of structure in the data between the most and the least coherent suggests that our measure of coherence does pick up for some unobserved heterogeneity between these two categories. We have formally tested for each wave the null hypothesis that the most and the least coherent share the same monotonic relationship, i.e. that the Spearman's rho between μ and σ is the same for the two groups. The null hypothesis is always rejected at the 1 percent significance level, highlighting that the relationship is not the same by level of coherence. The same analysis has been carried out with the seven alternative measures of coherence, and the results, presented in Table D1 of the online appendix, confirm this difference.

These preliminary results seem to support our dual system approach, which aims to separate the most corrupted data from the least ones. Section 4 will analyze in more depth the cross-

[10]Instead of a dot to describe their probability distribution, a few respondents have the parameters (μ, σ) of their probability distribution described by a cross in the panels of Figure 2. These are the few respondents who will be excluded from the regressions in Table 4-Section 4 because they are outliers.

[11]Note that there are two Spearman's rho figures at the bottom of each panel in Figure 2. The first (e.g., 0.60 in Panel (a)) is computed with all the respondents of the category of coherence and wave under consideration. The second (e.g., 0.72 in Panel (a)) is computed excluding the "outliers" (e.g., excluding from the calculation the 11 respondents described by a cross in Panel (a)). We always refer to the first of these Spearman correlations in the text.

section distribution of (μ, σ) conditional on our measure of coherence. Before that, the next Subsection checks to see if there is a relationship between our measure of coherence and various individual characteristics.

3.3 Who's who in term of coherence?

We expect the measure of coherence to be correlated with some individual characteristics related to cognitive abilities (e.g., education and income), given that System 2 reflects what psychologists label the "need for cognition", i.e. a respondent's tendency to enjoy thinking, and exposure to statistical thinking (Kahneman, 2003, p.1467). On the contrary, other individual characteristics that may not be related to cognitive ability (e.g., religion) should not be correlated with our measure of coherence. We also expect our measure of coherence to be independent of the period of interview. Cognitive abilities change only slowly over time, so unless the general population experienced a jump in education or financial literacy, our measure of coherence in comparable samples of the general population should be distributed in a similar fashion across waves.[12]

Table 3 proposes regressions in which the dependent variable is the measure of coherence d^{98}. Most of the regressors here are dummies reflecting various respondent characteristics: education, total income in the past 12 months before taxes, gender, religion and race.[13] We also include a quadratic in age, and, to see if d^{98} varies over time, wave dummies. Note that we would have liked to know if respondents hold assets, but the SEE does not provide such information. This is a serious shortcoming that limits the analysis.

The first three columns of Table 3 describe how d^{98} varies with individual characteristics

[12]Note that Table 2 already suggests that this is the case. Table 2 shows that the 113 respondents categorized as the most coherent in wave 12 are those for whom $d^{98} \in [0, 0.10]$, so the most coherent have a measure of coherence which is lower than the upper bound 0.10 in wave 12. This upper bound is broadly the same in wave 13 and 14 (0.09 and 0.10, respectively). The lower bounds of respondents categorized as the least coherent are also remarkably similar across waves, d^{98} being higher than 0.25, 0.24 and 0.25 in waves 12, 13 and 14 respectively.

[13]In contrast to Dominitz and Manski (2010, Table 4), there is no category "American Indian" in our Table. This due to small cell sizes in our sample (8 respondents out of 979). These are included in the base group category "Other", as is the case for the 15 respondents who refused to answer the question.

and over time using least squares. The characteristics predict little of the variation in the measure of coherence, the R^2 being between 0.03 and 0.04 depending on the specification. In fact the estimated coefficients (constant term excluded) are jointly significantly different from zero (at the five percent level) only when the set of dummies for the respondent's income is included in the specification (Columns [1]-[2]). In Column [1], only one estimated coefficient (apart from the constant) differs significantly from zero (at the five percent level): those who earn between \$50000 and \$60000 have a significantly lower d^{98} than the base group, i.e. those who earn less than \$10000. Excluding the education dummies in Column [2] reinforces the effect of income: those who earn between \$20000 and \$60000 are more coherent than the base group (the joint hypothesis that the four associated coefficients are equal is not rejected (P-value=0.69)). As expected, the race and religion dummies as well as the wave dummies always attract insignificant coefficients close to zero.

Note that the least squares estimates of Columns [1]-[3] might fail to account for some respondents having d^{98} levels which are limit values (i.e. zero or one).[14] Thus, Columns [4]-[6] of Table 3 provide Tobit regressions. The results are broadly the same, except that respondents with a MS or a PhD are now significantly more coherent when we do not control for income (Column [6]).

This analysis has also been carried out with the seven alternative measures of coherence. The results, presented in Tables E1-E7 of the online appendix, are qualitatively similar: the few variables significantly different from zero are mainly related to income and schooling. For some alternative measures of coherence, two other variables become significant: the gender dummy and the quadratic in age. Being older or being a man (slightly) reduces the measure of coherence in some cases. But, again, the individual characteristics predict little of the variation in the measures of coherence.

Remark that the analyses in Tables 3 and E1-E7 has been realized considering the final

[14]For instance, 77 respondents have a $d^{98} = 0$ and one has a $d^{98} = 1$ in the specification in Column [1]. The 829 other respondents have nonlimit values (strictly between zero and one).

sample of 979 respondents, those for whom we can find a unique fitted value for (μ, σ) (the numbers of observations in the regressions of Table 3 are slightly lower (907/960), given that some respondents did not answer some questions, in particular their level of education). However, and as mentioned in Subsection 3.1, d^{98} can be computed for more respondents (1125). If we consider this sample (Tables F1-F7 of the online appendix), we find similar results. The main difference concerns those with a BA/BS or a MS/PhD: they are significantly more coherent at the 1 percent significance level, whatever the method (least squares or Tobit) and the specification.[15]

Lastly, and interestingly, education appears to explain not only the degree of coherence but also nonresponse. Indeed, the nonrespondents, i.e. those who did not answer the preliminary questions or those who answered the preliminary questions but not the sequence of probabilistic questions, are overrepresented in the less educated respondents: DM (2010, p.10) report that 48 percent of those with no postsecondary education did not respond versus 18 percent of those with a bachelor's degree. We can thus safely conjecture that most of the nonrespondents are far from being coherent.

4 Interpreting expectations of equity returns

Section 3 has shown that there is a significant positive relationship between μ and σ for the most coherent respondents. On the contrary, we do not find any relationship for the least coherent. Following our dual system approach, we view the responses of the least coherent as noisy answers that are difficult to interpret. Hence, this Section mainly focuses on the 33 percent most coherent respondents of the final sample of 979 respondents. We first show in Subsection 4.1 that a simple linear model accurately describes the relation between risk and return for

[15] One can also be interested in the level of coherence of the 146 ($= 1125 - 979$) "special" respondents for whom we can compute a measure of coherence but cannot fit a subjective distribution. Most of these special respondents reported the same probability at the four thresholds and are far from being coherent. The (0.33, 0.66)-quantiles of d^{98} for the 146 "special" respondents are (0.145, 0.49), while the (0.33,0.66)-quantiles of d^{98} for the final sample of 979 respondents are (0.10,0.25) whatever the wave (see Table 2 and footnote 12).

these respondents, and propose an interpretation of this empirical regularity. Subsection 4.2 explores how expectations behave across the three waves of the SEE survey. This permits us to sketch a model of belief formation.

4.1 On the nature of the risk-return trade-off

The previous Section has shown that there is an increasing monotonic trend between μ and σ for the most coherent. Panels (a), (d) and (g) of Figure 2 also suggest that this monotonic relationship is linear for this category of individuals. To see whether this linear relation holds, we estimate the following equation (via least squares), for the $N_{j,t}$ respondents in category of coherence j interviewed at wave t:

$$\mu_{i,j,t} = \gamma_{j,t} + \beta_{j,t} \times \sigma_{i,j,t} + \epsilon_{i,j,t} \ , \ i = 1, \cdots , N_{j,t} \qquad (4)$$

where $\gamma_{j,t}$ and $\beta_{j,t}$ are the intercept and the slope coefficients for the sample of Coherent j interviewed at wave t. We are particularly interested in the R^2, given that it reflects the goodness of fit of the regression line. To take into account the influence of outliers (which might affect the R^2), we carry out a DFITS diagnostic test and exclude respondent i of category j interviewed at wave t if he has a cutoff value of $|DFITS_{i,j,t}| > 2\sqrt{\frac{2}{N_{j,t}}}$, as suggested by Belsley *et al.* (1980).

Columns [1]-[3] in Table 4 present the estimates for the separate regressions fitted for each category of coherence for wave 12. Column [4] presents the results when we pool the categories. Observations which are flagged by the DFITS cutoff criterion, and so are excluded in the separate regressions of Columns [1]-[3], are not included in the pooled regression. Note (iii.) at the bottom of the Table provides the results if we do not exclude these observations. The online appendix presents the results for the two other waves, as well as the results when respondents are ranked according to the other measures of coherence.

The coefficient of determination is 72% for the most coherent. This is noteworthy in individual cross-section data, and indicates that the linear model fits the data remarkably well. This contrasts with the least coherent sample, where the coefficient of determination is close to zero. The results for the waves 13 and 14, presented in Table A1 of the online appendix, confirm the goodness of fit of the linear model for the most coherent, the R^2 being 72% for wave 13 and 51% for wave 14. The analysis with the other measures of coherence produce similar results (Section B of the online appendix): the linear regression always explains a significant percentage of the total variation in μ for the most coherent.

A possible interpretation. The existence of a robust linear relation proves that the heterogeneity between the most coherent respondents takes a particular form. In what follows, we propose an interpretation of this fact. We assume that the most coherent respondents only differ in the way in which they interpret the mutual fund described in the SEE scenario ("*a type of mutual fund known as a diversified stock fund*"). This definition does not fully specify the composition of the portfolio, nor does it specify whether the fund will be rebalanced during the year under consideration. Respondents might have different portfolios in mind, ranging from relatively safe investments to much more risky ones. In this, faced with the same scenario, some might anticipate high returns, but high risk, while others anticipate low returns associated with low risk. At the extreme, those who have a safe portfolio in mind, should anticipate the market to perform in the same way as a risk-free asset. According to this interpretation, the parameters of Equation 4 are interpretable in financial terms: the intercept term $\gamma_{1,t}$ is the riskless rate at wave t, and the slope coefficient $\beta_{1,t}$ is the ex-ante price of risk for the sample of the most coherent interviewed at wave t.[16]

If this interpretation holds, the intercept term for the most coherent in wave t should not differ significantly from the risk-free rate during this period. As a proxy for the risk-free rate,

[16]In our interpretation, the most coherent respondents differ in how they assess the composition of the portfolio but share a common price for risk. Alternatively, we might consider that the composition of the portfolio is common knowledge and the most coherent respondents simply differ in their expectations. This alternative view cannot be rejected. But remark that it is hard to imagine that some might see the same investment as perfectly safe and others as extremely risky. Furthermore, this alternative interpretation does not allow us to interpret the parameters in a natural way, and precludes any comparison to observable economic variables.

we take the one-year U.S. LIBOR rate (RF_t).[17] During wave 12, this interest rate oscillated in the range $RF_{12} \in [0.056; 0.063]$.[18] The estimated riskless rate is $\hat{\gamma}_{1,12} = 0.059$ (Column [1] of Table 4), and its estimated standard error is $\hat{se}_{\hat{\gamma}_{1,12}} = 0.033$. Hence, the 95% confidence interval for $\gamma_{1,12}$ is $[-0.005, 0.123]$, and the null hypothesis $H_0 : \gamma_{1,12} = RF_{12}$ versus $\gamma_{1,12} \neq RF_{12}$ cannot be rejected, $\forall RF_{12} \in [0.056; 0.063]$. Table 5 gives the results for waves 13 and 14 (where the null hypothesis is not rejected either), as well as the results for the other measures of coherence. The estimated value $\hat{\gamma}_{1,t}$ is usually not significantly different from the risk-free rate in wave t, with one exception in wave 13 (when respondents are ranked according to d_{1000}) and two exceptions in wave 14 (when respondents are ranked according to d^{80} and d^{Kol}).

With respect to $\beta_{1,t}$, there is no simple way to test whether this slope coefficient reflects the ex-ante price of risk. The ex-ante price of risk is related to expected returns and perceived risk in the year ahead. We can test the null hypothesis that $\beta_{1,t}$ is equal to a price of risk computed with realized outcomes, i.e. an ex-post price of risk. But if we test this null hypothesis, we are not testing the hypothesis that $\beta_{1,t}$ is equal to the ex-ante price of risk; we are testing the hypothesis that the respondents classified as the most coherent have specific expectations.

To clearly understand the problem, let's consider wave 12 (July-November 1999). The estimate $\hat{\beta}_{1,12}$ varies over the range $[0.84; 1.34]$ depending on the measure of coherence considered (see Table 5). These estimated prices of risk appear very optimistic if we compare them to the historical mean and standard deviation of one-year equity returns in the U.S. DM (2010, p.17) report that *"the historical mean real return through the end of the twentieth century was about 0.07 and the standard deviation was near 0.20. The inflation rate was low and non-volatile during the sample period, generally being between 0.02 and 0.03."* Based on this evidence, DM consider that if individuals follow the long-run historical record of equity returns

[17] The U.S. LIBOR rate is a daily reference rate based on the interest rates at which banks borrow funds from other banks. Other proxies are possible, such as the FED Funds rate or the short-term U.S. T-Bill rate. Their values are however very similar to the U.S. LIBOR rate.

[18] The one-year U.S. LIBOR rate was 0.057 on July 1st, 1999 (when wave 12 began), and it was 0.062 on November 30th, 1999 (when wave 12 finished). During this period, it reached a minimum of 0.056 on July 21st, and a maximum of 0.063 on October 27th.

to predict future returns, the one-year-ahead return in nominal term should be around 0.095, and the standard-deviation around 0.18. If we consider that the risk-free rate was around 0.06 during wave 12, the price of risk should be around $0.20 (\simeq \frac{0.095-0.06}{0.18})$. Whatever the measure of coherence used to rank the respondents, $\hat{\beta}_{1,12}$ always differs significantly from this "long-run" price of risk. However, remark that wave 12 took place during a notable bull market, i.e. a boom in the stock market accompanied by widespread optimism. The Standard & Poor 500 (S&P 500) return was between 0.20 and 0.40 per year,[19] and therefore much higher than the historical average. With this information in mind, the range $\hat{\beta}_{1,12} \in [0.84; 1.34]$ appears plausible. [20]

4.2 Towards a model of expectations formation in financial markets

As discussed in the previous Subsection, it is impossible to test the null hypothesis that $\beta_{1,t}$ is equal to the ex-ante price of risk. At best, we can check if (i) the variation in this ex-ante price of risk across waves makes sense, and (ii) these movements allow us to discriminate between some classes of models of expectations formations.

The three waves of the SEE offer the opportunity to observe expectations during radically different periods: waves 12 and 13 (July-November 1999 and February-May 2000) took place when the return on the S&P 500 was above 0.20 per year as mentioned above, while wave 14 (September 2000-March 2001) took place at the beginning of a notable bear market, i.e. a steady drop in the stock market accompanied by widespread pessimism. At the beginning of September 2000 (when wave 14 began), the index was over 1500 points (and so was very close to its all-time intraday high of 1552 on March 24th, 2000). It lost more than 20% of its value

[19] The S&P 500 price returns per year were 0.31, 0.26 and 0.20 in the years 1997, 1998 and 1999, respectively. And DM (Table 1, p.29) report that the S&P 500 price return reached 0.38 the year before wave 12, i.e. from September 1998 to August 1999.

[20] Remark that this range of values [0.84;1.34] is also plausible if we compare it to the results in Lettau and Ludvigson (2010, Figure 11.2, p.673). Lettau and Ludvigson do not have expectational data. They use quarterly excess returns on the value-weighted stock market index from the Center for Research in Security Prices, and measure movements in the ex-ante expected return from period t to period $t+1$ (expected standard deviation of daily returns between t and $t+1$) as forecastable variation in the realized return (realized standard deviation), with e.g., the log consumption-wealth ratio in t as a forecasting variable. Over the period 1953-2001, they find that the price of risk was the highest during the quarters of 1996 (above 1.5), and the lowest during the last quarter of 2000 (around -0.4).

over seven months, bottoming out at under 1200 points at the end of March 2001 (when wave 14 finished).[21] Hence, we might expect the ex-ante price of risk in wave 14 to be lower than that in the other waves.

To evaluate the impact of this decline in the stock market on expectations, we compare the estimated ex-ante prices of risk across waves. Depending on the measure of coherence considered, $\hat{\beta}_{1,12}$ varies over the range $[0.84; 1.34]$; $\hat{\beta}_{1,13} \in [0.79; 1.18]$; $\hat{\beta}_{1,14} \in [0.75; 1.07]$ (see Table 5). For the eight measures of coherence, we test three hypotheses: (i) the null hypothesis $H_0 : \beta_{1,12} = \beta_{1,13}$ versus $H_1 : \beta_{1,12} \neq \beta_{1,13}$; (ii) $H_0 : \beta_{1,13} \leq \beta_{1,14}$ versus $H_1 : \beta_{1,13} > \beta_{1,14}$; and (iii) $H_0 : \beta_{1,12} \leq \beta_{1,14}$ versus $H_1 : \beta_{1,12} > \beta_{1,14}$. Table 5 indicates that the estimated ex-ante price of risk in wave 13 does not differ significantly from that in wave 12 (except when the respondents are preliminary ranked according to d_{1000}). The estimated ex-ante price of risk in wave 14 is significantly lower, although this result is not entirely robust (when the respondents are preliminary ranked according to $\overline{d}$ and d^{Kol}). Furthermore, even when the estimated ex-ante price of risk in wave 14 is statistically lower than in the other waves, the size of the gap may look small: for instance, when respondents are ranked according to d^{98}, the practical difference between the point estimates $\hat{\beta}_{1,12}$ and $\hat{\beta}_{1,14}$ is equal to 0.26(=1.29-1.02). The S&P 500 return was above 20% during waves 12 and 13 and lost more than 20% of its value during wave 14, so we may have expected a difference substantially higher than 0.26. Our results then suggests that the ex-ante price of risk is weakly procyclical, i.e. the ex-ante price of risk falls somewhat as stock market returns drop. This result is in line with Hudomiet *et al.* (2010), who analyze similar data from the Health and Retirement Study (HRS) over the period February 2008-February 2009 (i.e. before and after the crash of October 2008), and find that expectations reacted slowly to the crash.

The procyclical behavior of the ex-ante price of risk can help us to discriminate between some classes of models of expectations formation. There are various ways of modeling expectations

[21] This bear market is considered to have ended in October 2002. The S&P 500 reached a low of 768 intraday on October 10th, 2002, losing 50% of its value between September 2000 and October 2002.

formation.[22] This section does not try to review all of them. At best, we try to discriminate between three classes of simple models, borrowed from DM. These three classes of models have normal subjective distributions and use the changes in the S&P 500 index to form the mean and variance of their subjective distributions. Note that we also assume that all respondents use a similar model, also we do not consider more sophisticated approaches in which respondents use mixtures of these models. The first class of models is a random-walk class which postulates that the changes in the stock market are best predicted by a normal distribution with mean and variance equal to their long term value. The second class of models are persistent, i.e. recent stock-market performance is a good predictor of future market performance. The last class of models considers that the stock market exhibits mean-reversal, i.e. high recent returns implies a return to the long-term mean in the near future.

Each class of models implies some restriction on the mean and the variance of expected returns. The results obtained for the most coherent allow us to rule out the random-walk and mean-reversal classes of models. The peculiarity of random walk models is to produce very similar predictions at each point of time, as long as inflation is low and non-volatile and the risk-free rate is stable. As we noted above, the estimated ex-ante price of risk is much higher than the long run price of risk. This rules out the possibility of simple random-walk models. Mean-reversal models predict that higher returns, like those observed when wave 12 took place, should lead to lower returns in the near future. If this were the case, the probability of lower returns should have increased in wave 13, after an additional year of historically high returns. However, the estimated ex-ante price of risk in wave 13 does not differ significantly from that in wave 12. In contrast, the persistence model seems to do a decent job at organizing the data. The fact that the estimated price of risk fell when stock market returns dropped in wave 14 means that respondents consider that recent stock market performance will persist. We would like to be more precise about what "recent" means. But when defining a persistence model,

[22]See Pesaran and Weale (2006) for an exhaustive review of these alternative models.

a respondent might use, e.g., the last year's performance on the S&P 500 or only that in the last three months to form (μ_i, σ_i). So the best that we can say is that the information used probably covers more than the last few months, given that the estimated price of risk in wave 14 is lower but does not collapse either. Nevertheless, this remains a fuzzy calibration.

5 Conclusion

Taking into account the pathologies which affect answers in surveys, this paper has used the specific design of the SEE survey to construct several measures of coherence based on a principle of dual reasoning *à la* Kahneman. Our main contribution is to show that there is a highly significant linear relationship between expected returns and perceived risk for the most coherent respondents. On the contrary, the cross-section distribution of the least coherent respondents exhibits no particular pattern,as if they provided nothing but noise in their answers. This result is robust across waves and to the various measures of coherence. We view this result as supporting our dual system approach which aims to distinguish data by its level of corruption.

There was no particular reason to expect that a simple linear relation between risk and return would fit the data for the most coherent respondents. We propose a particular interpretation (respondents have different portfolios in mind) which allows us to interpret the parameters of the estimated model as the risk-free rate and the price of risk. The empirical evidence does not reject this interpretation.

Last, the price of risk for the most coherent appears to be procyclical, as it is lower when individuals were surveyed during a period of widespread pessimism. This suggests that the expectations of coherent respondents can accurately be described by a simple persistence model, as we have defined it.

This last finding is, however, only suggestive. With some measures of coherence there is no evidence of procyclicality; in addition, even when we find procyclicality the time-series dimension is too short to consider this result as definitive. Further progress in understanding how people form expectations of equity returns requires longitudinal data. Some surveys are longitudinal, but have their own specific disadvantages. For instance, the Michigan survey, also analyzed by DM (2010), is longitudinal, but does not enable us to estimate subjective distributions (it only asks one probabilistic expectation question regarding positive equity returns). Another longitudinal survey is the HRS. Like the SEE, the HRS allows us to estimate subjective distributions in some waves, but it only interviews individuals over the age of 50. It is also important to investigate how expectations of equity returns affect individual's behavior regarding asset holdings. But the SEE does not ask respondents if they hold stocks. Last, it would be interesting to know if our measure of coherence is correlated with measures of cognitive capacities. We have particularly in mind questions which measure the ability to perform basic numerical operations, such as those asked in the Survey of Health, Ageing and Retirement in Europe (SHARE). Christelis *et al.* (2010) use SHARE data to show that this ability is strongly correlated with the propensity to invest in stocks.

References

BELSLEY, D. A., KUH, E. and WELSH, R. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: Wiley.

BOSSAERTS, P. (2002), *The Paradox of Asset Pricing*, Princeton University.

CHRISTELIS, D., JAPPELLI, T. and PADULA, M. (2010), "Cognitive Abilities and Portfolio Choices", *European Economic Review*, vol. 54 nº 1: pp. 18–38.

DOMINITZ, J. (2001), "Estimation of Income Expectations Models Using Expectations and Realization Data", *Journal of Econometrics*, vol. 102 nº 2: pp. 165–195.

DOMINITZ, J. and MANSKI, C. F. (1997), "Using Expectations Data To Study Subjective Income Expectations", *Journal of the American Statistical Association*, vol. 92 nº 439: pp. 855–867.

DOMINITZ, J. and MANSKI, C. F. (2004), "The Survey of Economic Expectations", Technical report, http://faculty.wcas.northwestern.edu/~cfm754/see_introduction.pdf.

DOMINITZ, J. and MANSKI, C. F. (2007), "Expected Equity Returns and Portfolio Choice: Evidence from the Health and Retirement Study", *Journal of the European Economic Association*, vol. 5 nº 2-3: pp. 369–379.

DOMINITZ, J. and MANSKI, C. F. (2010), "Measuring and Interpreting Expectations of Equity Returns", *Journal of Applied Econometrics*, forthcoming.

FLACHAIRE, E. and HOLLARD, G. (2008), "Individual Sensitivity to Framing Effects", *Journal of Economic Behavior & Organization*, vol. 67 nº 1: pp. 296–307.

HUDOMIET, P., KEZDI, G. and WILLIS, R. J. (2010), "Stock Market Crash and Expectations of American Households", Technical report, university of Michigan.

Hurd, M. D. (1999), "Anchoring and Acquiescence Bias in Measuring Assets in Household Surveys", *Journal of Risk and Uncertainty*, vol. 19 n° 1-3: pp. 111–136.

Kahneman, D. (2003), "Maps of Bounded Rationality: Psychology for Behavioral Economics", *American Economic Review*, vol. 93 n° 5: pp. 1449–1475.

Kahneman, D. and Frederick, S. (2005), "A Model of Heuristic Judgement", in Holyoak, K. J. and Morrison, R. G. (editors), *The Cambridge Handbook of Thinking and Reasoning*, Cambridge Handbook in Psychology, pp. 267–294.

Lettau, M. and Ludvigson, S. C. (2010), "Measuring and Modelling Variation in the Risk-Return Trade-off", in Ait-Sahalia, Y. and Hansen, L. P. (editors), *Handbook of Financial Econometrics*, Elsevier, vol. 1, chapter 11, pp. 617–690.

Manski, C. F. (2004), "Measuring Expectations", *Econometrica*, vol. 72 n° 5: pp. 1329–1376.

Pesaran, M. H. and Weale, M. (2006), "Survey Expectations", in Elliott, G., Granger, C. and Timmermann, A. (editors), *Handbook of Economic Forecasting*, Elsevier, vol. 1, chapter 14, pp. 715–776.

Tversky, A. and Kahneman, D. (1974), "Judgment under Uncertainty: Heuristics and Biases", *Science*, vol. 185: pp. 1124–1131.

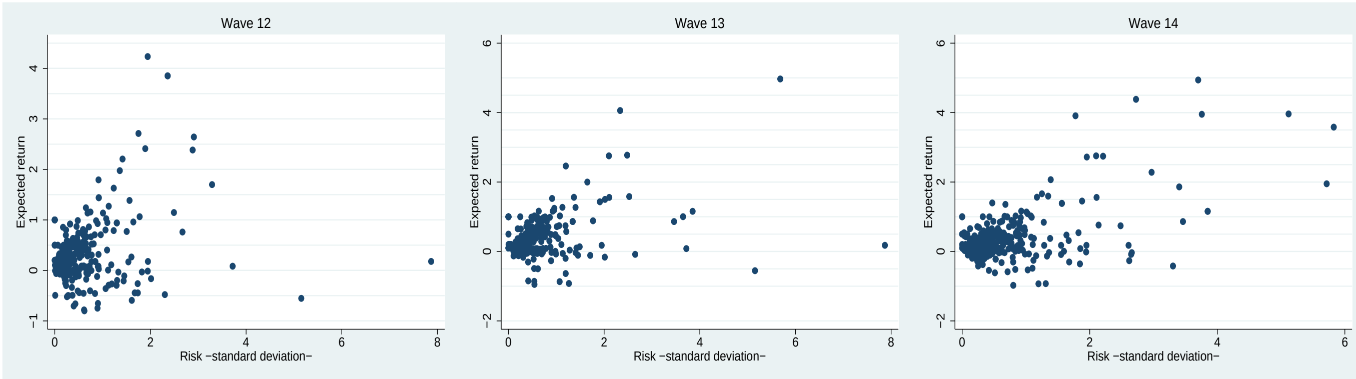
Van Rooij, M., Lusardi, A. and Alessie, R. (2007), "Financial Literacy and Stock Market Participation", Research paper 162, Michigan Retirement Research Center.

Zimmer, A. C. (1984), "A Model for the Interpretation of Verbal Predictions", *International Journal of Man-Machine Studies*, vol. 20: pp. 121–134.

Table 1: Fitted subjective distributions of mutual fund returns

		Mean	Std. Dev.	Quantile 0.25	0.50	0.75
Wave 12	N=340					
	μ	0.32	0.60	0.03	0.17	0.50
	σ	0.57	0.74	0.18	0.34	0.67
Wave 13	N=264					
	μ	0.40	0.63	0.08	0.28	0.60
	σ	0.67	0.87	0.22	0.43	0.76
Wave 14	N=375					
	μ	0.36	0.69	0.03	0.20	0.47
	σ	0.65	0.81	0.19	0.39	0.79
Total	N=979					
	μ	0.35	0.64	0.04	0.20	0.50
	σ	0.63	0.80	0.19	0.39	0.74

Note: i. For the sake of comparison, DM (2010, Table 4, p.32) use a sample of 986 respondents. The (0.25, 0.50, 0.75)-quantiles of μ in their sample are (0.04, 0.20, 0.50), and are thus identical to the (0.25, 0.50, 0.75)-quantiles of μ in our sample. The same holds for their (0.25, 0.50, 0.75)-quantiles of σ, which are (0.19, 0.39, 0.74) in their sample as well as in our sample. Last, note that the means/standard deviations of μ and σ are 0.36/0.65 and 0.63/0.80 in their sample, respectively. These values are broadly similar to ours.



(a) (b) (c)

Figure 1: Plots of the expected return (μ) against risk (σ) by wave

Table 2: Number of individuals by type of coherence and wave

	Wave 12	Wave 13	Wave 14	Total
Coherent 1	113 [0, 0.10]	88 [0,0.09]	125 [0,0.10]	326
Coherent 2	114 [0.10, 0.25]	88 [0.09, 0.24]	125 [0.10,0.25]	327
Coherent 3	113 [0.25,0.95]	88 [0.24,1]	125 [0.25,0.9]	326
Total	340	264	375	979

Notes: i. The first line in each cell shows the number of respondents in the category in question.
ii. The bottom entries are distance intervals. For instance, the 113 respondents classified as Coherent 1 in the wave 12 are individuals for whom $d^{98} \in [0, 0.10]$.

Figure 2: Plot of expected return (μ) against risk (σ) by type of coherence and by wave

Notes: i. A small number of respondents have the parameters (μ, σ) of their probability distribution depicted by a cross instead of a dot in each figure. These respondents correspond to those who will be excluded from the regressions in the columns of Table 4 because they are outliers (see Section 4 for more details).

ii. At the bottom of each panel, we report two Spearman rank correlation coefficients between μ and σ. One rho is computed including all the respondents of the category of coherence and wave. The other rho is computed excluding the outliers (see Note (i)). The superscripts $*$, $**$ and $***$ represent 10, 5 and 1% significance of these rhos, respectively.

Table 3: Who's who in term of coherence (d^{98})

		OLS			Tobit		
		[1] d^{98}	[2] d^{98}	[3] d^{98}	[4] d^{98}	[5] d^{98}	[6] d^{98}
Education							
	≤Grade 11		Base group			Base group	
	High school graduate	-0.006		-0.011	-0.000		-0.005
		(0.056)		(0.056)	(0.046)		(0.046)
	Attended college	-0.039		-0.047	-0.041		-0.050
		(0.051)		(0.051)	(0.041)		(0.041)
	Associated degree	-0.049		-0.059	-0.051		-0.061
		(0.055)		(0.055)	(0.047)		(0.047)
	BA/BS	-0.058		-0.067	-0.057		-0.065
		(0.052)		(0.051)	(0.041)		(0.041)
	Masters or PhD	-0.075		-0.085	-0.070		-0.080*
		(0.053)		(0.052)	(0.044)		(0.044)
	Professional	-0.029		-0.028	-0.025		-0.024
		(0.071)		(0.070)	(0.058)		(0.058)
Income							
	Income<10000		Base group			Base group	
	10000≤Income<20000	-0.003	-0.027		-0.011	-0.036	
		(0.034)	(0.033)		(0.032)	(0.032)	
	20000≤Income<30000	-0.035	-0.056*		-0.041	-0.062**	
		(0.029)	(0.030)		(0.031)	(0.031)	
	30000≤Income<40000	-0.038	-0.063**		-0.051	-0.076**	
		(0.030)	(0.030)		(0.032)	(0.031)	
	40000≤Income<50000	-0.023	-0.057*		-0.029	-0.064**	
		(0.031)	(0.031)		(0.033)	(0.032)	
	50000≤Income<60000	-0.061**	-0.084***		-0.069*	-0.091**	
		(0.029)	(0.031)		(0.037)	(0.036)	
	Income≥60000	0.008	-0.02		0.008	-0.026	
		(0.028)	(0.028)		(0.030)	(0.029)	
	Did not know/refused	-0.032	-0.062**		-0.035	-0.065**	
		(0.029)	(0.028)		(0.031)	(0.030)	
Age							
	Age/100	-0.399	-0.549**	-0.414	-0.436*	-0.588**	-0.448*
		(0.28)	(0.26)	(0.27)	(0.26)	(0.25)	(0.25)
	$(Age/100)^2$	0.004	0.006**	0.004	0.004*	0.006**	0.005*
		(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)
Gender							
	Male	-0.017	-0.012	-0.019	-0.015	-0.009	-0.016
		(0.013)	(0.013)	(0.013)	(0.014)	(0.014)	(0.014)
Race							
	White	-0.014	-0.020	-0.0092	-0.011	-0.018	-0.006
		(0.025)	(0.025)	(0.025)	(0.026)	(0.026)	(0.026)
	Black	0.015	0.036	0.014	0.017	0.039	0.016
		(0.037)	(0.039)	(0.037)	(0.038)	(0.037)	(0.038)
	Asian	0.030	0.015	0.040	0.035	0.021	0.047
		(0.045)	(0.046)	(0.045)	(0.048)	(0.049)	(0.049)
	Other		Base group			Base group	
Religion							
	No religion	0.010	0.018	0.012	0.012	0.020	0.014
		(0.025)	(0.024)	(0.025)	(0.027)	(0.027)	(0.027)
	Roman Catholic	0.023	0.035	0.021	0.023	0.035	0.020
		(0.023)	(0.023)	(0.023)	(0.026)	(0.026)	(0.026)
	Protestant	-0.0026	0.009	-0.006	-0.006	0.006	-0.009
		(0.022)	(0.021)	(0.022)	(0.025)	(0.025)	(0.025)
	Christian	-0.009	-0.000	-0.011	-0.010	-0.004	-0.014
		(0.022)	(0.022)	(0.023)	(0.026)	(0.026)	(0.026)
	Other		Base group			Base group	
Wave dummies							
	Wave 12		Base group			Base group	
	Wave 13	0.000	-0.003	0.000	-0.001	-0.004	-0.001
		(0.016)	(0.017)	(0.017)	(0.018)	(0.018)	(0.018)
	Wave 14	0.007	-0.002	0.008	0.007	-0.000	0.009
		(0.015)	(0.015)	(0.015)	(0.017)	(0.016)	(0.017)
Constant		0.368***	0.377***	0.357***	0.374***	0.384***	0.357***
		(0.073)	(0.062)	(0.073)	(0.071)	(0.063)	(0.069)
Observations		907	960	907	907	960	907
R^2		0.04	0.03	0.03			
Significance of the regression: P-value		0.03**	0.02**	0.13	0.10*	0.03**	0.22

Notes: i. $*$, $**$ and $***$ refer to significance at the 10, 5 and 1% levels respectively.
ii. Heteroskedasticity-robust standard errors in parentheses.
iii. "Significance of the regression" shows the P-value of the joint test of the hypotheses that all the coefficients except the constant are zero. For least-squares estimates this corresponds to the P-value associated with the F-statistic. For Tobit estimates the P-value is that associated with the likelihood-ratio statistic which has a limiting chi-squared distribution under the null hypothesis.

Table 4: Least squares regressions of μ on σ by type of coherence - wave 12-

	[1]	[2]	[3]	[4]
	Coherent 1	Coherent 2	Coherent 3	All types
	μ	μ	μ	μ
Intercept	0.059*	0.011	-0.061*	0.153***
	(0.033)	(0.018)	(0.035)	(0.038)
σ	1.290***	0.797***	-0.011	0.186*
	(0.099)	(0.034)	(0.043)	(0.10)
R^2	0.72	0.79	0.00	0.04
Observations	102	108	107	317
(Observations excluded)	11	6	6	23

Notes: i. $*$, $**$ and $***$ refer to significance at the 10, 5 and 1% levels respectively.
ii. Heteroskedasticity-robust standard errors in parentheses.
iii. If we do not exclude the observations that are flagged by the DFITS cutoff criterion, we obtain:

$$\mu = 0.07 + 1.41^{***}\sigma + \hat{\epsilon} \text{ for Coherent 1 } (R^2 = 0.77 \text{ and } N = 113)$$

$$\mu = -0.00 + 0.79^{***}\sigma + \hat{\epsilon} \text{ for Coherent 2 } (R^2 = 0.68 \text{ and } N = 114)$$

$$\mu = -0.08^{*} + 0.06\sigma + \hat{\epsilon} \text{ for Coherent 3 } (R^2 = 0.03 \text{ and } N = 113)$$

$$\mu = 0.18^{***} + 0.24^{**}\sigma + \hat{\epsilon} \text{ all types considered } (R^2 = 0.08 \text{ and } N = 340)$$

Table 5: Summary results for the most coherent for all measures of coherence

	Measure of coherence used to rank respondents							
	d^{98}	d^{90}	d^{80}	α	d_{1000}	α_{1000}	$\overline{d}$	d_{Kol}
$\hat{\beta}_{1,12}$	1.29***	1.29***	1.18***	1.04***	1.34***	1.06***	1.04***	0.84***
$\hat{\beta}_{1,13}$	1.18***	1.12***	1.01***	1.11***	0.79***	1.04***	0.93***	0.83***
$\hat{\beta}_{1,14}$	1.02***	0.80***	0.75***	0.76***	1.07***	0.85***	0.99***	1.01***
t-values of the test:								
$H_0 : \beta_{1,12} = \beta_{1,13}$ versus $H_1 : \beta_{1,12} \neq \beta_{1,13}$	0.87	1.33	1.61	-0.58	3.09***	0.17	0.56	0.1
$H_0 : \beta_{1,13} = \beta_{1,14}$ versus $H_1 : \beta_{1,13} > \beta_{1,14}$	1.21	2.3**	2.07**	5.34***	-1.67	1.58*	-0.37	-1.06
$H_0 : \beta_{1,12} = \beta_{1,14}$ versus $H_1 : \beta_{1,12} > \beta_{1,14}$	1.83**	3.15***	3.13***	2.50***	2.75***	1.42*	0.31	-1.09
$\hat{\gamma}_{1,12}$	0.06	0.06	0.06	0.07	0.02	0.04	0.09	0.03
95% ci	[-0.00;0.12]	[-0.00;0.12]	[0.00;0.12]	[0.00;0.13]	[-0.04;0.09]	[-0.03;0.12]	[-0.00;0.18]	[-0.04;0.11]
Reject $H_0 : \gamma_{1,12} = RF_{12}$?	No	No	No	No	No	No	No	No
$\hat{\gamma}_{1,13}$	0.07	0.07	0.08	0.07	0.24	0.09	0.06	0.03
95% ci	[-0.02;0.13]	[0.01;0.13]	[0.02;0.14]	[0.02;0.12]	[0.10;0.37]	[0.03;0.14]	[-0.04;0.16]	[-0.04;0.10]
Reject $H_0 : \gamma_{1,13} = RF_{13}$?	No	No	No	No	Yes	No	No	No
$\hat{\gamma}_{1,14}$	0.09	0.14	0.15	0.11	0.09	0.09	-0.01	-0.11
95% ci	[0.01;0.17]	[0.06;0.22]	[0.07;0.22]	[0.06;0.16]	[0.02;0.15]	[0.03;0.16]	[-0.07;0.06]	[-0.21;-0.02]
Reject $H_0 : \gamma_{1,14} = RF_{14}$?	No	Yes/No	Yes	Yes/No	No	No	Yes/No	Yes

Notes: i. $*$, $**$ and $***$ refer to significance at the 10, 5 and 1% levels respectively.
ii. RF_t, $t = 12, 13, 14$, stands for the risk-free rate in wave t, which is proxied by the daily one-year U.S. LIBOR rate. During wave 12 (July 1999-November 1999), this reached a minimum of 0.056 on July 21st, and a maximum of 0.063 on October 27th. So RF_{12} moved in the range $RF_{12} \in [0.056; 0.063]$. During wave 13 (February-May 2000), this interest rate reached a minimum of 0.067 (February 28th) and a maximum of 0.075 (May 31st), so $RF_{13} \in [0.067; 0.075]$. During wave 14 (September 2000-March 2001), RF_{14} reached a minimum of 0.045 (March 22nd, 2001) and a maximum of 0.069 (September 1st, 2000), so $RF_{14} \in [0.045; 0.069]$.
iii. "95% ci" stands for the 95% confidence interval. We provide such confidence intervals for $\gamma_{1,t}$, $t = 12, 13, 14$, to test $H_0 : \gamma_{1,t} = RF_t$. As an example, the 95% ci is [-0.00;0.12] for $\gamma_{1,12}$ when respondents have been preliminary ranked according to d^{98}. During wave 12, the one-year U.S. LIBOR oscillated in the range $RF_{12} \in [0.056; 0.063]$ (see Note ii. of this Table). Hence the null hypothesis $H_0 : \gamma_{1,12} = RF_{12}$ versus $\gamma_{1,12} \neq RF_{12}$ cannot be rejected, $\forall RF_{12} \in [0.056; 0.063]$.
iv. "Reject $H_0 : \gamma_{1,t} = RF_t$?" provides three possible results: "No", "Yes" and "Yes/No". "No" is when the null hypothesis cannot be rejected, whatever the value taken by RF_t during wave t. "Yes" is when the null hypothesis is rejected, whatever the value taken by RF_t during wave t. (For instance when respondents are ranked according to d^{80}, the 95% ci for $\gamma_{1,14}$ is [0.07;0.22]; $RF_{14} \in [0.045; 0.069]$ always falls outside of the 95% confidence interval.) Last, "Yes/No" is when the null hypothesis is rejected for some values taken by RF_t during wave t, but not rejected for the other values (for instance,when respondents are ranked according to α, the 95% ci for $\gamma_{1,14}$ is [0.06;0.16]; $RF_{14} \in [0.045; 0.069]$ falls outside of the 95% confidence interval for some values (those between 0.045 and 0.06) and falls inside for the other values (those between 0.06 and 0.069)).